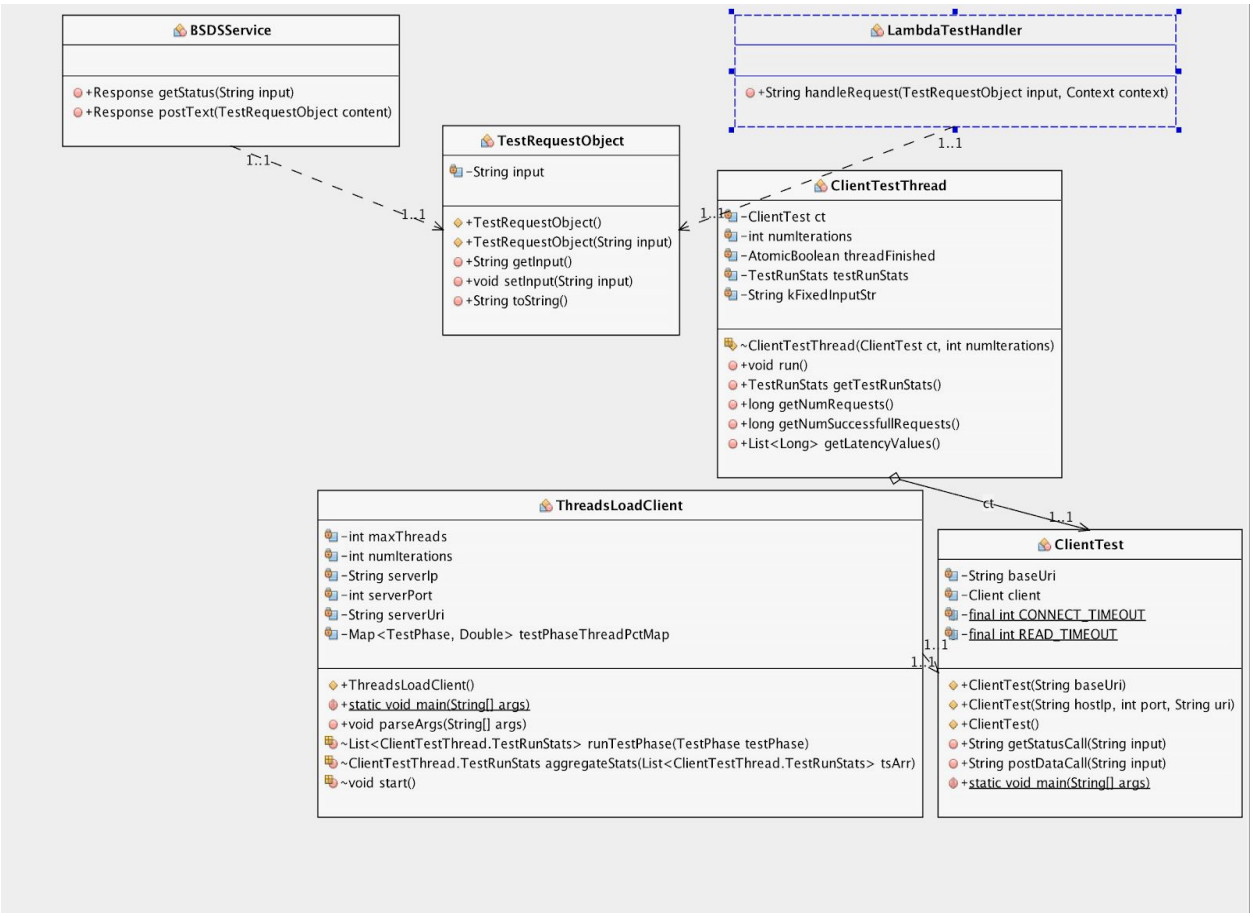
BSDS ASSIGNMENT 1

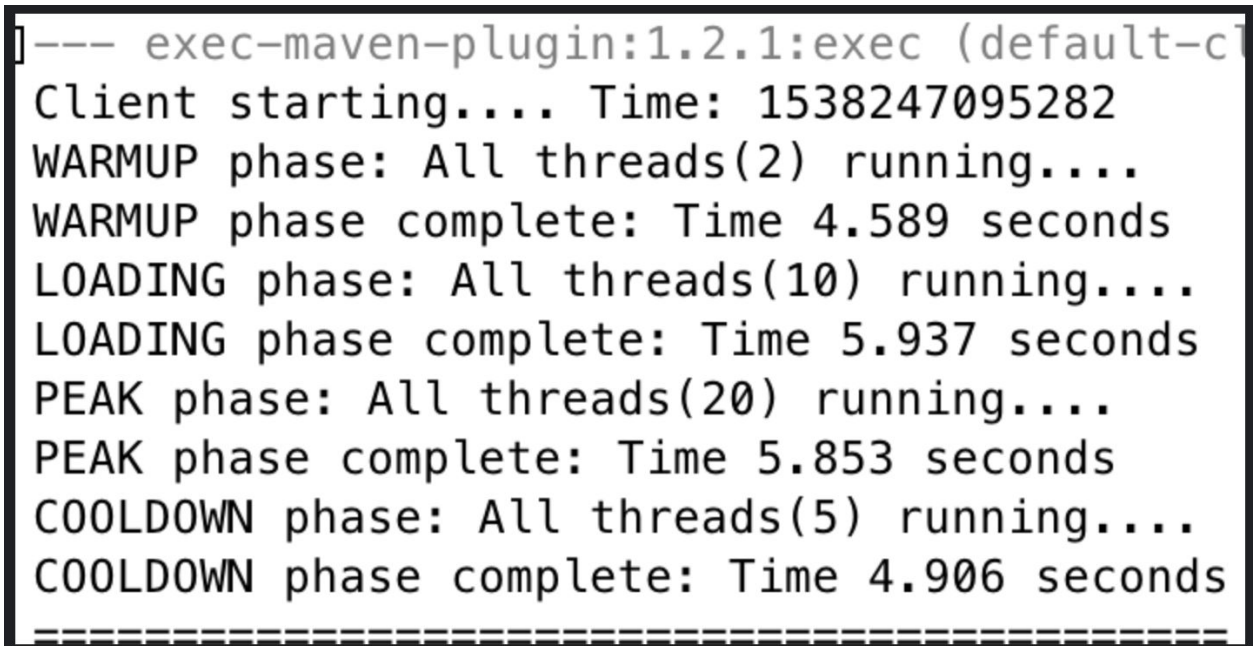
STEP 1: A 1 page overview of your design
Block Diagram

- BSDSService is simply a server class that has get and post methods.
- ThreadsLoadClient is class in which all the threads and iterations that we receive through command line is joined and run together in all four phases(Peak, Warmup, Loading).
- In LambdaTestHandler, lambda function for Lambda server is used.
- I used args4j for parsing command line.

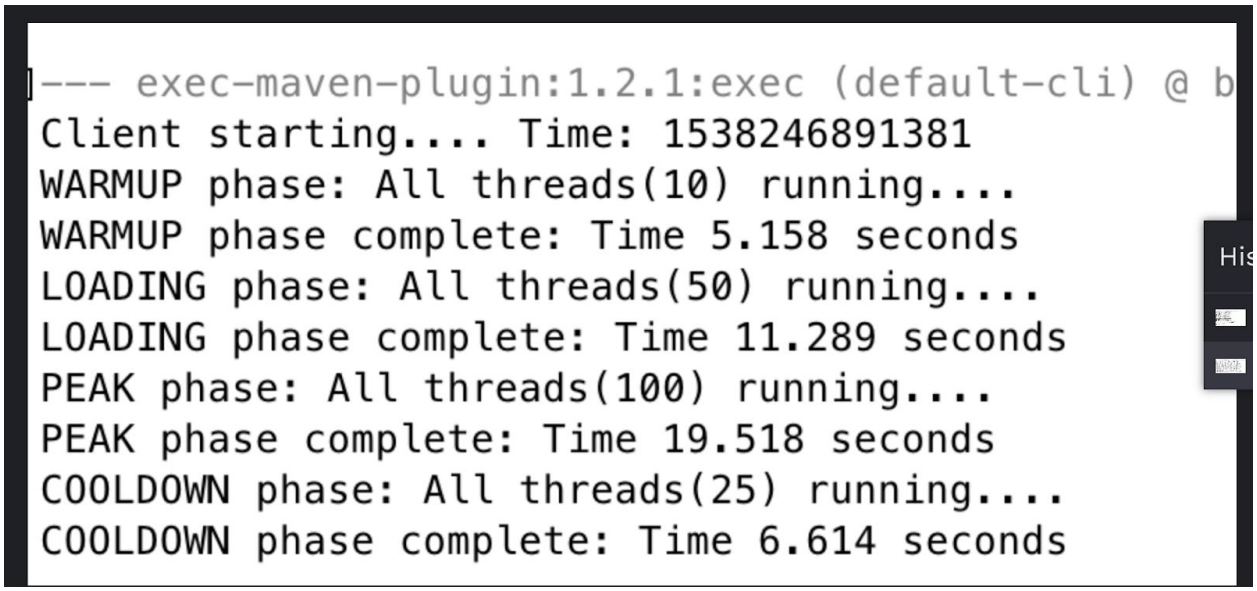


Step 2 : URL to your git repo.
https://github.com/manika0407/bsdscourse

Step 3 : Two screenshots for step 4
20 Threads



100 Threads



Step 4 : Two screenshots for step 5 (EC2 Instance)

EC2 20 threads 100 iterations Screenshot

```
]--- exec-maven-plugin:1.2.1:exec (default-cli) @ bsdsassignment ---
 Client starting.... Time: 1538247095282
 WARMUP phase: All threads(2) running....
 WARMUP phase complete: Time 4.589 seconds
 LOADING phase: All threads(10) running....
 LOADING phase complete: Time 5.937 seconds
 PEAK phase: All threads(20) running....
 PEAK phase complete: Time 5.853 seconds
 COOLDOWN phase: All threads(5) running....
 COOLDOWN phase complete: Time 4.906 seconds
 ========================================
 Total number of requests sent: 7400
 Total number of Successful responses: 7400
 Test Wall Time: 21.781 seconds
 Overall throughput across all phases: 339.74564987833435 rps.
 Mean Latency for all requests = 27.510852567567568 ms.
 Median Latency = 25.167 ms.
 P95 Latency = 42.65235 ms.
 P99 Latency = 67.07838000000001 ms.
```

EC2 100 Threads 100 Iterations Screenshot

```
 ------------------------------------------------------------------------

]--- exec-maven-plugin:1.2.1:exec (default-cli) @ bsdsassignment ---
 Client starting.... Time: 1538246891381
 WARMUP phase: All threads(10) running....
 WARMUP phase complete: Time 5.158 seconds
 LOADING phase: All threads(50) running....
 LOADING phase complete: Time 11.289 seconds
 PEAK phase: All threads(100) running....
 PEAK phase complete: Time 19.518 seconds
 COOLDOWN phase: All threads(25) running....
 COOLDOWN phase complete: Time 6.614 seconds
 ========================================
 Total number of requests sent: 37000
 Total number of Successful responses: 37000
 Test Wall Time: 43.06 seconds
 Overall throughput across all phases: 859.2661402693915 rps.
 Mean Latency for all requests = 69.3171265945946 ms.
 Median Latency = 47.2275 ms.
 P95 Latency = 171.6947 ms.
 P99 Latency = 556.5666 ms.
 ------------------------------------------------------------------------
```

Step 5 : Two screenshots for step 6 (against Lambda Server)

Lambda 20 Threads 100 Iterations

```
Client starting.... Time: 1538251380779
WARMUP phase: All threads(2) running....
WARMUP phase complete: Time 11.703 seconds
LOADING phase: All threads(10) running....
LOADING phase complete: Time 11.607 seconds
PEAK phase: All threads(20) running....
PEAK phase complete: Time 14.348 seconds
COOLDOWN phase: All threads(5) running....
COOLDOWN phase complete: Time 7.902 seconds
==========================================
Total number of requests sent: 7400
Total number of Successful responses: 7400
Test Wall Time: 46.071 seconds
Overall throughput across all phases: 160.6216491936359 rps.
Mean Latency for all requests = 56.60378527027027 ms.
Median Latency = 43.2325 ms.
P95 Latency = 111.04910000000001 ms.
P99 Latency = 318.88107 ms.
------------------------------------------------------------------
```

Lambda 100 Threads 100 Iterations

```
--- exec-maven-plugin:1.2.1:exec (default-cli) @ bsdsassignment ---
Client starting.... Time: 1538251743815
WARMUP phase: All threads(10) running....
WARMUP phase complete: Time 9.342 seconds
LOADING phase: All threads(50) running....
LOADING phase complete: Time 10.49 seconds
PEAK phase: All threads(100) running....
PEAK phase complete: Time 28.281 seconds
COOLDOWN phase: All threads(25) running....
COOLDOWN phase complete: Time 8.782 seconds
==========================================
Total number of requests sent: 37000
Total number of Successful responses: 37000
Test Wall Time: 57.38 seconds
Overall throughput across all phases: 644.8239804810038 rps.
Mean Latency for all requests = 81.19252210810811 ms.
Median Latency = 40.529 ms.
P95 Latency = 160.87125 ms.
P99 Latency = 1206.11639 ms.
------------------------------------------------------------------
```

Step 7: Stress testing

I was initially checking my ec2 and lambda for just 20/100 and 100/100 as stated in assignment but i eventually tried to change the thread count and iteration count one by one. When I increased the thread count to 110 , 120, 130 etc , they were working fine but for 150 thread-count, it was initially breaking and I was receiving connection timed out exception.
And then I changed the following two things-

1. Increased the client connection timeout to server and read timeout for requests to 5 seconds.
2. Increased the number of threads in server by changing tomcat server configurations to 500.

And after making these two changes, I got all the requests accepted for 150/150 and 200/200. Following I have submitted the screenshot for 200/200

```
Client starting.... Time: 1538259193704
WARMUP phase: All threads(20) running....
WARMUP phase complete: Time 11.182 seconds
LOADING phase: All threads(100) running....
LOADING phase complete: Time 39.037 seconds
PEAK phase: All threads(200) running....
PEAK phase complete: Time 81.014 seconds
COOLDOWN phase: All threads(50) running....
COOLDOWN phase complete: Time 20.352 seconds
==========================================
Total number of requests sent: 148000
Total number of Successful responses: 147988
Test Wall Time: 151.917 seconds
Overall throughput across all phases: 974.2161838372269 rps.
Mean Latency for all requests = 133.13293114306566 ms.
Median Latency = 61.8585 ms.
P95 Latency = 592.7460500000001 ms.
P99 Latency = 1108.13205 ms.
```

WHAT BROKE IT ?

It was working fine for 100/ 150/200/500 threads but as the number of threads were increasing, the number of failed requests was increasing as well.
When I checked for 1000 threads, the number of failed requests was 5%(3700000-3548107) out of total sent requests. (3700000).

HOW?

Increasing client side timeouts and number of threads did initially help but they also significant increase in p99 latencies. Because of multiple requests to the server and network layer is not able to process them fast enough, there are lot of packet retransmits which overall increases latencies and timeouts. Cpu (35% server), memory(10%) and network bandwidth (20-30MB/s) on both client and server side were well within limits. So it indicated towards network level congestion and retransmits. I increased kernel socket read/write buffers from 4k to 10's of MB and kernel max connection limit to 2k both on client/server but it didn't help that much as retransmits and timeouts were still high. Multiple clients from different network devices can still push the server more so this maybe some sort of overload/qos backoff/congestion etc. from client side wifi router or other hops b/w source and destination...

SCREENSHOT TO SHOW EXECUTION OF 1000 THREADS AND 1000 ITERATIONS

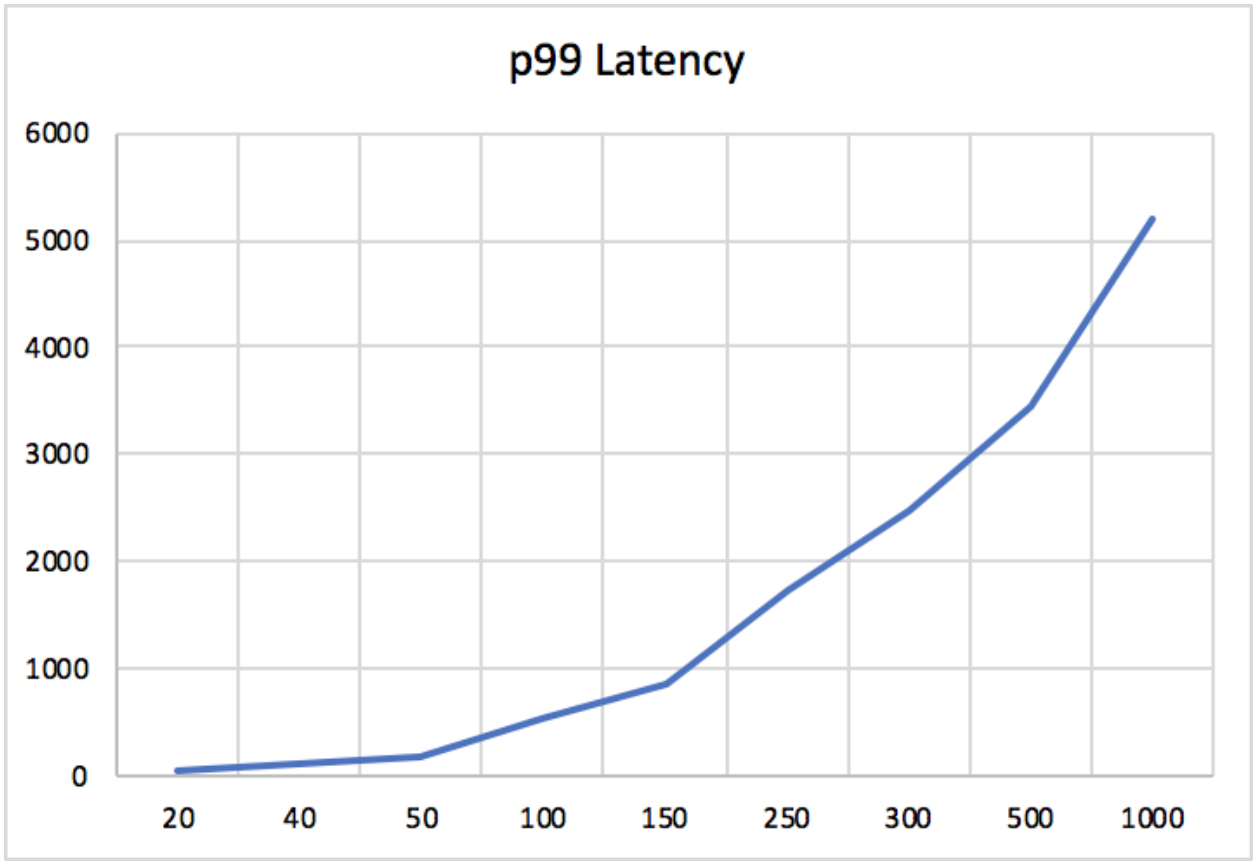
```
Building bsdsassignment 1.0-SNAPSHOT
------------------------------------------------------------------------

--- exec-maven-plugin:1.2.1:exec (default-cli) @ bsdsassignment ---
Client starting.... Time: 1538261612440
WARMUP phase: All threads(100) running....
WARMUP phase complete: Time 200.707 seconds
LOADING phase: All threads(500) running....
LOADING phase complete: Time 987.475 seconds
PEAK phase: All threads(1000) running....
PEAK phase complete: Time 1918.751 seconds
COOLDOWN phase: All threads(250) running....
COOLDOWN phase complete: Time 474.04 seconds
==========================================
Total number of requests sent: 3700000
Total number of Successful responses: 3548107
Test Wall Time: 3581.317 seconds
Overall throughput across all phases: 1033.139484720286 rps.
Mean Latency for all requests = 457.2120171141964 ms.
Median Latency = 72.346 ms.
P95 Latency = 2387.5108999999998 ms.
P99 Latency = 5201.43332 ms.
------------------------------------------------------------------------
```

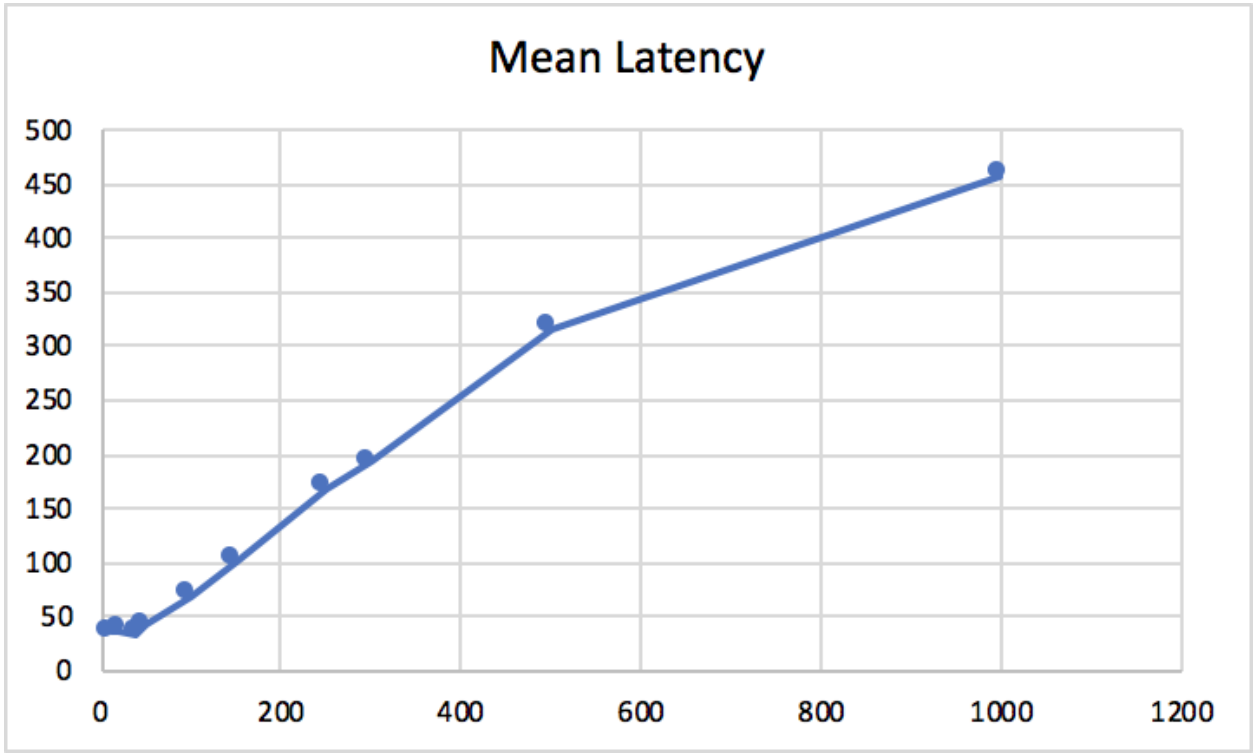
Step 7: CHARTING

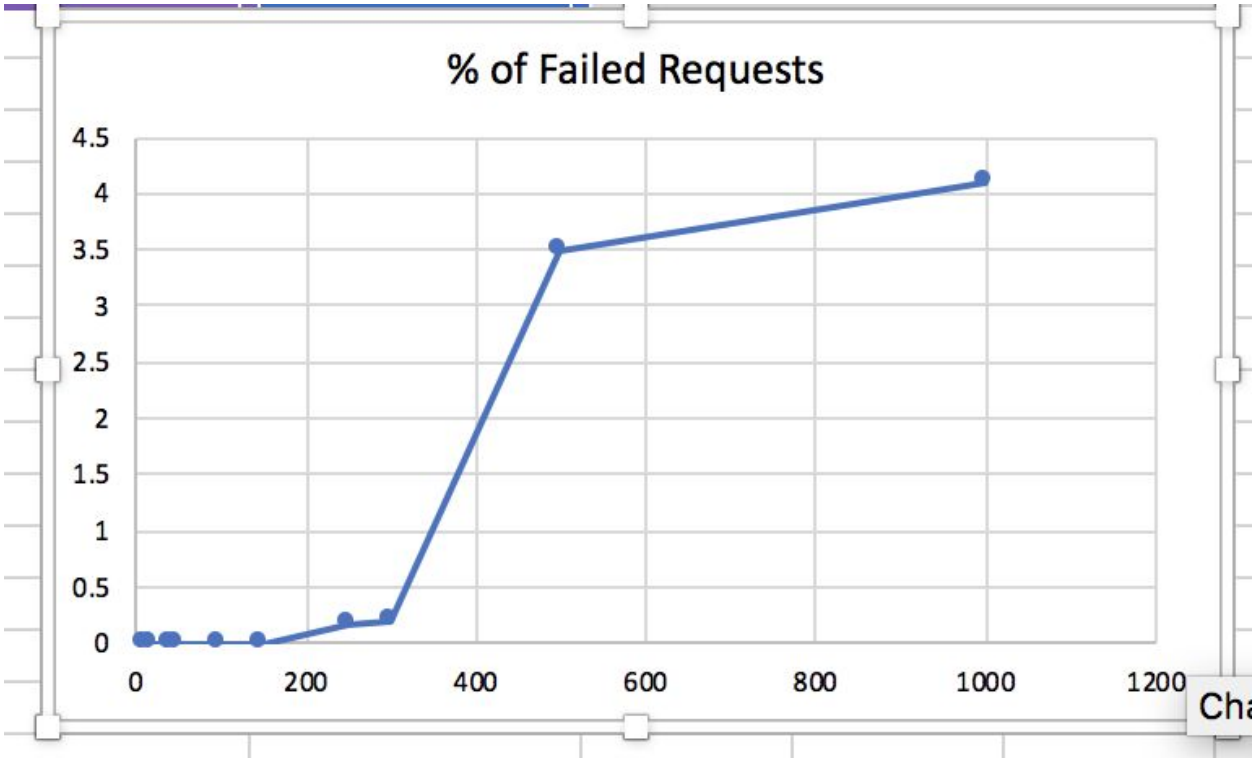
- For completing the charting, the measurements that I used are mean, p99 latency and % of failed requests. And I printed the results of p99 and mean against number of threads.
- For all the graphs, I plotted the number of threads on x-axis and p99, mean and % of failed requests.

Number of Threads against P99 Latency (in milliseconds).



Graph of Number of Threads against Mean Latency




% of Failed Requests
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
200
400
600
800
1000
1200